Journal of Biomedical Informatics 41 (2008) 944–952

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Single and multiple time-point prediction models in kidney transplant outcomes

Ray S. Lin [a,*], Susan D. Horn [b,c], John F. Hurdle [c], Alexander S. Goldfarb-Rumyantzev [d]

[a] Biomedical Informatics, Stanford University, MSOB X-215, 251 Campus Drive, Stanford, CA 94305-5479, USA
[b] Institute for Clinical Outcomes Research, Salt Lake City, UT, USA
[c] Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[d] Division of Nephrology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Article history:
Received 22 September 2007
Available online 22 March 2008

Keywords:
Survival analysis
Kidney transplantation
Graft survival
Recipient survival
Logistic regression
Cox proportional hazards models
Artificial neural networks

ABSTRACT

This study predicted graft and recipient survival in kidney transplantation based on the USRDS dataset by regression models and artificial neural networks (ANNs). We examined single time-point models (logistic regression and single-output ANNs) versus multiple time-point models (Cox models and multiple-output ANNs). These models in general achieved good prediction discrimination (AUC up to 0.82) and model calibration. This study found that: (1) Single time-point and multiple time-point models can achieve comparable AUC, except for multiple-output ANNs, which may perform poorly when a large proportion of observations are censored, (2) Logistic regression is able to achieve comparable performance as ANNs if there are no strong interactions or non-linear relationships among the predictors and the outcomes, (3) Time-varying effects must be modeled explicitly in Cox models when predictors have significantly different effects on short-term versus long-term survival, and (4) Appropriate baseline survivor function should be specified for Cox models to achieve good model calibration, especially when clinical decision support is designed to provide exact predicted survival rates.

1. Introduction

Prognosis prediction of medical treatments is a clinically important yet challenging problem. Prediction of survival before treatment potentially could facilitate patients' decision making and improve survival by altering clinical practice. Individual factors predicting survival have been studied extensively in various clinical domains, but the complex interaction of these factors makes outcome prediction a significant challenge. Statistical and machine learning models, such as regression models or artificial neural networks, could be developed based on pre-treatment variables and be used to identify patients who may not benefit by the treatment or to optimize the modifiable factors in the treatment (e.g., treatment parameters) in order to achieve the best predicted outcome.

Survival can be modeled as survival time (continuous outcomes) or survival rates at specific time points (dichotomous outcomes). For predicting dichotomous outcomes, two categories of prediction models, single versus multiple time-point models, have been reviewed and compared in the literature [1]. The single time-point model can predict the survival rate at only one specific time point. Typical examples of single time-point models are logistic regression [2] and single-output artificial neural networks (ANNs) [1]. When predicting the survival rates at multiple time points, multiple independent models need to be aggregated: each model is developed independently and then used to predict the survival rate for one time point.

Alternatively, a multiple time-point model can generate a continuous survival curve across all time points and can predict any time-specific survival rate. Examples of this type of model include Cox proportional hazards models [3] and multiple-output ANNs [1]. Some researchers have found that multiple time-point models achieve better performance in survival prediction than the aggregation of single time-point models [1].

Another frequent challenge of aggregated single time-point models is non-monotonic prediction: a model may predict a lower survival rate for a specific patient at one time point, yet another model predicts a higher survival rate for the same patient at a later time point. This limitation is inevitable in the aggregation of single time-point models, as all the models are mutually independent and there is no mechanism to synchronize the predicted survival rates across the models. Non-monotonic prediction is spurious and difficult to interpret clinically. As a model for facilitating clinical decision making, non-monotonic prediction should be minimized as much as possible.

This paper compared single and multiple time-point models in the prediction of graft and recipient survival in kidney transplantation. This study investigated two regression modeling techniques and two types of ANNs: logistic regression and single-output ANNs (as single time-point models) versus Cox models and multiple-output ANNs (as multiple time-point models). A retrospective analysis

* Corresponding author. Fax: +1 650 725 7944.
E-mail address: raylin@stanford.edu (R.S. Lin).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.03.005

of United States Renal Data System (USRDS) was performed. Models predicting graft and recipient survival were constructed based on pre-transplant variables. Prediction discrimination, model calibration, and percentage of non-monotonic prediction were examined. The advantages and limitations of each type of model are discussed and summarized. These findings are generalizable to survival prediction in different clinical domains.

2. Background

2.1. Logistic regression

Logistic regression models the relationship between a dichotomous outcome (e.g., survival or failure) and the predictors. It assumes that the mean of the outcome is linearly related to the predictors.

Suppose x is a vector of predictors, and p is the response probability to be modeled. Logistic regression has the form:

$$\text{logit}(p) \equiv \log\left(\frac{p}{1-p}\right) = \alpha + \beta' \cdot x \tag{1}$$

where α is the intercept parameter, β is the vector of slope parameters, and β' is the transpose of β.

Logistic regression is not designed to handle censored data. However, by constructing multiple logistic regression models at different points along the survival time (e.g., 1-yr, 3-yr, 5-yr), censored data can be utilized in at least some of the models instead of being discarded completely. For example, if a record is censored at the 4th year, it can be utilized in models for predicting 1-yr and 3-yr but not 5-yr survival.

Logistic regression commonly is used to explain the effect of predictors on an outcome as well as to produce patient-specific survival rates.

2.2. Cox model

The Cox proportional hazards model [3] is a semi-parametric method. It can handle censored data and thus is widely used in survival analysis to explain the effect of predictors on outcomes. Cox models assume a parametric form for the effects of the predictors but allow an unspecified form for the underlying survivor function. The survival time of each patient is assumed to follow the hazard function, $h_i(t)$, expressed as

$$h_i(t) = h(t; z_i) = h_0(t)\exp(z_i' \cdot \beta(t)) \tag{2}$$

where $h_0(t)$ is an arbitrary and unspecified baseline hazard function, z_i is the vector of measured predictors for the i-th individual, and $\beta(t)$ is the vector of unknown regression parameters associated with the predictors. The vector $\beta(t)$ is a function of time and is assumed to be the same for all individuals. When $\beta(t)$ is constrained to be constant over time, the effects of the predictors (i.e., the hazard ratio $exp(z_j' \cdot \beta(t))$ are assumed to be the same for short-term and long-term survival. This is known as the *proportional hazard assumption*. The proportional hazard assumption can be tested by global goodness-of-fit statistics [4] or chi-square statistics of Schoenfeld residuals [5]. If the proportional hazard assumption does not hold, non-proportional hazards need to be modeled by allowing $\beta(t)$ changing over time to accommodate the *time-varying effects* of the predictors (i.e., the predictors can have different effects on short-term versus long-term survival) [5].

The survivor function can be expressed as

$$S(t; z_i) = \exp\left(-\int_0^t h_i(u)\mathrm{d}u\right) = \exp\left(-\int_0^t h_0(u)\exp(z_i' \cdot \beta(u))\mathrm{d}u\right) \tag{3}$$

$$S_0(t) = \exp\left(-\int_0^t h_0(u)\mathrm{d}u\right) \tag{4}$$

where $S_0(t)$ is the baseline survivor function.

To predict a patient-specific survival rate, the baseline survivor function has to be specified either based on empirical data derived from product-limit estimators (i.e., Kaplan–Meier survival curves [6]) or by parametric models (e.g., Weibull model [1]).

2.3. Logistic regression and Cox models in clinical prediction

Logistic regression and Cox proportional hazard models have been used widely in survival analysis, including investigating the effect of individual predictors on survival and constructing models to predict patient-specific survival rates based on the interaction of all the predictors in various clinical domains [1]. In kidney transplantation, both methods have been used to identify risk factors of recipient and graft survival [7]; however, only a few models were developed to predict patient-specific survival rates after transplantation [8].

Logistic regression and Cox models in general assume independence of the predictors. They were not designed to handle complex interactions among predictors and are not often used to model non-linear relationships among predictors and outcomes.

2.4. Artificial neural networks

Artificial neural networks (ANNs) have been used to model complex and non-linear functions since they were introduced by computer scientists in artificial intelligence [9]. ANNs consist of a densely interconnected set of units. Each unit takes a number of real-valued inputs and produces a single real-valued output, which may serve as the input of other units [9]. These units are organized into several layers (Fig. 1). Each unit in the first layer (called *input layer*) takes the value from one predictor; and each unit in the last layer (called *output layer*) produces the prediction for one of the outcomes. There may be layers in between (called *hidden layers*), which calculate the weighted sum of inputs from the previous layer and produce the output for the next layer by applying a transformation function (e.g., pure linear or logistic sigmoid transformation) to the weighted sum. The outcome vector o can be represented as a function of the vector of predictors x:

$$o = f_2(w_2' \cdot f_1(w_1' \cdot x)) \tag{5}$$

where w_1 is the vector of weights associated with the outputs of the input layer with f_1 as the corresponding transformation function, and w_2 is the vector of weights associated with the outputs of the hidden layer with f_2 as the corresponding transformation function. The weights are adjusted based on the training data in order to minimize an error estimate function, such as mean squared error or

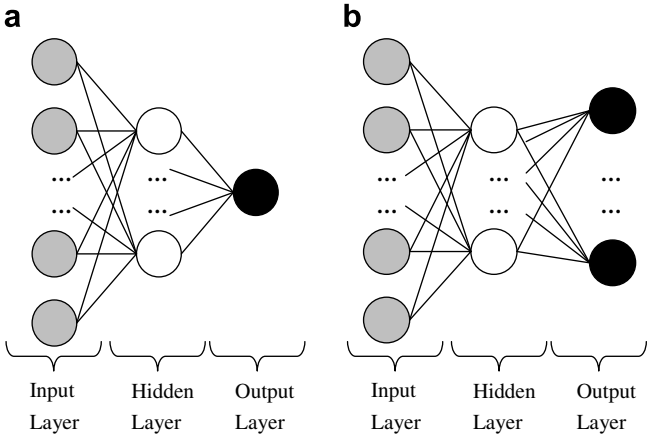

Fig. 1. Single-output and multiple-output ANNs.

cross-entropy error. Different algorithms for minimization have been developed (e.g., see [9,10]).

An ANN is identical to logistic regression if it consists of only one input layer and one output layer and uses a logistic sigmoid transformation function [11]. However, an ANN can model more complex relationships among the predictors and outcomes than logistic regression if it contains one or more hidden layers. In each hidden layer, the predictors are linearly combined, transformed, and mapped into a different vector space, which allows the network to represent complex interactions among the predictors and to model non-linear relationships.

An ANN is able to model more than one outcome at the same time. A single-output ANN produces one output and models one outcome (Fig. 1a) whereas a multiple-output ANN can model more than one outcome at the same time (Fig. 1b). Studies have suggested that a set of single output ANNs may perform similar to the corresponding multiple output ANN [1,12].

2.5. ANNs in clinical prediction

Prediction models based on ANNs have been explored in survival analysis in various clinical domains [13]. However, only few models were built to predict outcomes of kidney transplantation [14–16]. ANN models are rarely used to identify the effect of individual predictors. This is mainly because it is in general very difficult to interpret the models in terms of the effect of individual predictors [1], regardless of how precisely the overall models can predict the outcomes.

2.6. Performance comparison

Performance of prediction models based on ANNs and regression techniques has been studied and compared in various clinical outcomes [17]. A methodology review of logistic regression and ANNs shows that out of 72 studies, ANNs performed better in 51% of the studies while logistic regression performed better in 7%, and they showed no significant difference in 42% of the studies [11]. In the comparison of Cox models and ANNs, ANNs were found to achieve better prediction in certain domains [18,19], but other studies demonstrated similar performance between ANNs and Cox models [17].

The advantages and limitations of ANNs and regression models have been contrasted and summarized elsewhere [1,11]. However, these models have never been compared in a clinical dataset from the perspective of single versus multiple time-point models.

3. Materials and methods

This study investigated logistic regression, Cox models, single-output, and multiple-output ANNs in predicting kidney transplantation outcomes. These models were compared and discussed for the first time from the perspective of single versus multiple time-point models based on a clinical dataset.

3.1. Dataset

The dataset was derived from USRDS, 2003 version, which collects clinical and demographic data on patients with end-stage renal disease (ESRD), as well as from the United Network for Organ Sharing (UNOS), which collects data on transplant outcomes.

The study included recipients who underwent transplantation during the period beginning January 1, 1995 and ending December 31, 2002. Records with missing information on outcomes or model predictors were excluded. In total, 57,389 unique recipients were identified.

3.2. Outcome variables

Two outcome variables were analyzed in this study: (1) graft survival (time between the transplantation and allograft failure or censor) and (2) recipient survival (time between the transplantation and recipient death or censor).

3.3. Censoring

Recipient survival status was censored at the earliest time of lost to follow-up or study completion date. The graft survival status was censored at the earliest time of lost to follow-up, study completion date, or the date of recipient death if the recipient died with a functioning graft. The numbers of censored, survival, and failure observations for four time points, 1-yr, 3-yr, 5-yr, and 7-yr, are summarized in Table 1.

3.4. Predictors

The initial set of predictors for survival were selected from pre-transplant variables based on literature review [20,21] and our previous research [8]. Different subsets of predictors were included in the models to examine the predictive power of each predictor. The final set of predictors is the following:

(1) *Recipient variables:* Age; gender; race; height; weight; cause of ESRD; history of hypertension, diabetes, or cardiovascular disease; duration between date of current transplantation and failure date of the previous transplantation (if applicable); dialysis modality prior to the current transplant (hemodialysis, peritoneal dialysis, or none); predominant ESRD service (hemodialysis, peritoneal dialysis, transplantation, or none); and primary source of pay for treatment (used as a surrogate for socioeconomic status).
(2) *Donor variables:* Donor type (living or cadaveric, heart-beating or not); age; gender; race; height; weight; and cause of death.
(3) *Transplantation parameters:* Number of matched HLA antigens; cold storage time; and procedure type.

3.5. Data cleaning

Erroneous values for the predictors were eliminated. For example, the valid ranges for donors' and recipients' heights and weights were based on the United States CDC Growth charts for those younger than 18 years of age. For those 18 or older, heights and weights were based on acceptable ranges: height (122 to 274 cm), weight (23 to 180 kg).

3.6. Prediction models

3.6.1. Single time-point models

The study investigated two types of single time-point models: logistic regression and single-output ANNs. Both types of models

Table 1
Numbers of observations for survival variables at different time points of follow-up

Time point	Censored	Survival	Failure
Graft survival at different time points of follow-up			
1-yr	10,975	42,883	3531
3-yr	26,879	24,832	5678
5-yr	39,533	10,641	7215
7-yr	47,139	2495	7755
Recipient survival at different time points of follow-up			
1-yr	8463	46,032	2894
3-yr	23,927	28,192	5270
5-yr	37,051	13,168	7170
7-yr	45,846	3447	8096

were constructed for predicting 1-yr, 3-yr, 5-yr, and 7-yr survival of graft survival and recipient survival.

Since a single time-point model predicts the survival rate at only one time point, each of the four time points was analyzed by a separate model. In total, eight logistic regression models and eight single-output ANN models were constructed for the four time points predicting the two survival variables.

Each ANN model consists of three layers. There were 71 units in the input layer (corresponding to the 71 model predictors), 1 unit in the output layer (corresponding to the outcome), and 10 units in the hidden layer. To determine the number of units in the hidden layer, models with 140, 70, 50, 40, 30, 20, 10, 5 units in the hidden layers were examined. The model with 10 units in the hidden layer was selected because it achieved the best performance with shortest training time. Different transformation functions (logistic sigmoid, tangent sigmoid, saturated linear, and pure linear) were evaluated, and logistic sigmoid achieved the best prediction discrimination (defined in Sec. 3.7) and was selected. All the predictors were scaled to the interval between 0 and 1: continuous predictors were converted using linear transformation, and categorical predictors were treated with dummy coding. To minimize the error function, two algorithms were explored: scaled conjugate gradient [10], and gradient descent with momentum and adaptive learning rate [9]. The former achieved much better performance and thus was selected. Early-stopping of the training [9] was performed to avoid overfitting via monitoring the model performance in a holdout validation set of 20% of the training cases.

Neither logistic regression or ANN models are able to handle censored data completely. Therefore, only non-censored observations (Table 1) were analyzed in the models.

3.6.2. Multiple time-point models

Two types of multiple time-point models were investigated in this study: Cox models and multiple-output ANNs. In contrast to single time-point models, these two types of models are able to predict survival rates at multiple time points.

Cox models predict continuous survival curves over time. The 1-yr, 3-yr, 5-yr, and 7-yr survival rates were derived from the predicted survival curve. Two different Cox models were constructed to predict the two different survival variables. The baseline survivor functions were derived empirically as the average survival of all the cases in the dataset based on Kaplan–Meier survival curves [6]. Since Cox models can account for censored data, the complete dataset (i.e., 57,389 observations) was analyzed.

In the first step of model development, proportional hazards were assumed; that is, the effects of the predictors were assumed to be constant over time (no time-varying effects). The proportional hazard assumption was tested by global goodness-of-fit statistics [4], and the predictors that violate proportional hazard assumption were then identified by Schoenfeld residuals. In the second step, time-interaction terms were added into the Cox models to capture the time-varying effects of the predictors that violated proportional hazard assumption [5]. Global goodness-of-fit was tested again to assure validity of the models. The Cox models developed in the first step (with no time-varying effects) and those developed in the second step (with time-varying effects) were compared and discussed.

Similarly, two multiple-output ANN models were developed for the two survival variables. Each model predicted survival rates at all four time points. The ANN models consisted of 71 units in the input layer, 40 units in the hidden layer, and 4 units in the output layer. Transformation function was logistic sigmoid, and the algorithm for minimization was scaled conjugate gradient [10]. Decisions were made based on experiments similar to those described in single-output ANN models. Since multiple-output ANNs are not able to make use of an observation if the observation is censored at *any* of the time points being analyzed, only the observations that were not censored at the 7-yr were included in the models.

3.7. Performance assessment

Performance of the models was assessed by prediction discrimination, model calibration, and percentage of non-monotonic predictions.

1. Prediction discrimination was measured by area under the receiver operating characteristic (ROC) curve (AUC) [22] or equivalently, the *c* statistics [11].
2. Model calibration was measured by Hosmer–Lemeshow goodness-of-fit test. This test evaluates the degree of correspondence between the predicted survival rates and the observed survival over groups spanning the entire range of predicted rates. All observations were sorted by predicted survival rates and then divided into 10 groups. The difference of expected and observed survival in each group was calculated based on Chi-Square statistics with eight degrees of freedom. A Chi-Square value smaller than 15.51 corresponds to a *p*-value larger than 0.05, which indicates that there was no statistically significant difference between the predicted and observed survival rates [23].
3. The percentage of non-monotonic predictions was measured based on the number of unique recipients who have at least one non-monotonic prediction divided by the number of the total recipients.

Tenfold cross-validation [9] was conducted. The whole dataset was randomly divided into 10 groups and the following procedure was repeated for 10 iterations: in the *k*-th iteration, group *k* was left as the testing set and the models were trained based on the remaining 9 groups. AUC measurements and Hosmer–Lemeshow goodness-of-fit tests were applied to the prediction of the survival rates for observations in the testing set.

Logistic regression and Cox models were developed using SAS statistical analysis software, version 9.1.3. The models were constructed by the LOGISTIC and the PHREG procedures while the baseline survivor functions were derived from Kaplan–Meier curves by the LIFETEST procedure. All the ANN models were constructed by Neural Network Toolbox of MATLAB, version 7.0.4.

4. Results

Cox models were first developed assuming no time-varying effects. Global goodness-of fit statistics showed the proportional hazard assumption was violated in these models. Predictors violating this assumption were identified using Schoenfeld residuals; they were recipient age, recipient race, duration between date of current transplantation and failure date of the previous transplantation, predominant ESRD service, primary source of pay for treatment, and procedure type. Cox models were then rebuilt to accommodate time-varying effects of these predictors.

The AUC performance measure is presented in Table 2. Since the evaluation was done based on tenfold cross-validation, both the mean and the standard deviation are shown. The AUC for logistic regression varied from 0.71 to 0.81, which indicated fair to good performance. The AUC for single-output ANNs were 1% to 2% higher than logistic regression in both graft and recipient survival across all the four time points. Prediction discrimination in Cox models was significantly improved with the introduction of time-varying effects.

In comparison to single time-point models, Cox models (with time-varying effects) showed comparable AUC with logistic regression and single-output ANNs, while multiple-output ANNs performed similarly to the single time-point models in the 7-yr survival prediction but much worse than the other three types of models in the 1-yr prediction. The difference was particularly substantial in predicting 1-yr and 3-yr recipient survival, where the other three types of models performed similarly.

The Hosmer–Lemeshow Chi-Square values are shown in Table 3. The values for logistic regression and single-output ANNs in both outcomes were all below 15.51, indicating predicted survival rates were not statistically different from observed rates at significance level $\alpha = 0.05$. Model calibration for logistic regression and single-output ANNs are almost equal. However, Cox models showed poor model calibration in all predictions. Multiple-output ANNs achieved moderate to good model calibration only in 7-yr survival but much worse than Cox models at the other three time points.

Table 4 shows the percentage of non-monotonic prediction in logistic regression and ANN models. Logistic regression produced very little non-monotonic prediction in both graft survival and recipient survival (0.52% and 0.24%, respectively). Non-monotonicity was more common in ANNs. Single-output ANNs produced 2.34% and 2.81% non-monotonic predictions while the percentages for multiple-output ANNs reached 5.46% and 8.40% in the two survival variables. Measurement for non-monotonicity was not done for the Cox models since they will not produce a non-monotonic prediction as long as the baseline survivor functions are monotonic (which they were).

The 71 predictors in the models were categorized as significant predictors if the *p*-value of the Wald Chi-square statistic was smaller than 0.05 in logistic regression; otherwise, they were categorized as non-significant predictors. Table 5 shows the agreement of the predictors' effects in 1-yr versus 7-yr models. In predicting graft survival, 11 out of 71 predictors were significant in the 1-yr model but not significant in the 7-yr model (e.g., donor type and cold storage time); 10 were significant in the 7-yr model but not significant in the 1-yr model (e.g., recipients' history of diabetes and cause of ESRD). Similar results were found in predicting recipient survival: 8 predictors were significant in the 1-yr model but not significant in the 7-yr model (e.g., donor cause of death and cold storage time); 16 were significant in the 7-yr model but not significant in the 1-yr model (e.g., recipients' cause of ESRD and the duration between the date of current transplantation and the failure date of the previous transplantation).

5. Discussion

In survival analysis, it is commonly believed that Cox models are more appropriate than logistic regression [1,24]. One of the major advantages of Cox models is the ability to make the best use of censored observations. This increases the sample size and thus improves performance of the model substantially since real-world survival data are often highly censored. Furthermore, since Cox models produce continuous survival curves across all time points, the prediction is comprehensive and monotonic over time. In contrast, logistic regression can generate non-monotonic prediction since they are an aggregation of several independent models. A similar analogy may also be applied to ANN models. Multiple-output ANNs are more comprehensive as they predict survival rates at more than one time point whereas single-output ANNs may suffer from more serious non-monotonic prediction.

5.1. Non-monotonic prediction

This study found that non-monotonic predictions of logistic regression models were fairly rare in the USRDS dataset. Less than 1% of the recipients had non-monotonic predicted survival curves based on logistic regression. Single-output ANNs generated more non-monotonic predictions. Surprisingly, the non-monotonicity did not decrease but rather increased up to three times in multiple-output ANNs. No evidence was found that there is an internal mechanism in multiple-output ANNs for synchronizing the predicted survival rates across different time points. Using multiple-output ANNs did not provide an advantage of eliminating non-monotonic prediction. Another topology of ANNs, the sequential ANNs, were proposed and found to reduce non-monotonicity in survival prediction [1]. Future study is needed to examine the performance of sequential ANNs in this dataset.

5.2. Prediction discrimination

The study found that single time-point and multiple time-point models in general achieved comparable AUC at all time points, except for multiple-output ANNs, whose performance was significantly lower than the other three types of models at 1-yr, 3-yr, and 5-yr predictions. Comparing regression models and ANNs, single-output ANNs performed similar to logistic regression and Cox models (with time-varying effects). Similar results were found in previous literature [11,15,18,19]. Single-output ANNs did not have a substantial advantage over logistic regression, which indicates there might not be strong interaction between predictors in the dataset, and relationships between outcomes and individual predictors were relatively linear.

Multiple-output ANNs performed comparable to logistic regression, Cox models (with time-varying effects), and single-output ANNs in 7-yr prediction for both graft and recipient survival, but the performance decreased drastically in 1-yr and 3-yr prediction and was the worst among all the models. This is probably due to this model excluding an observation if it was censored at any of the time points being analyzed. In survival data, an observation

Table 2
Area under the ROC curve (AUC)

Time point	Logistic regression	Single-output ANN	Cox model* (no time-varying effects)	Cox model (with time-varying effects)	Multiple-output ANN
Graft survival					
1-yr	0.71 ± 0.01	0.73 ± 0.01	0.65 ± 0.01	0.72 ± 0.01	0.61 ± 0.01
3-yr	0.72 ± 0.01	0.74 ± 0.01	0.67 ± 0.01	0.73 ± 0.01	0.68 ± 0.01
5-yr	0.75 ± 0.01	0.77 ± 0.01	0.71 ± 0.01	0.74 ± 0.01	0.73 ± 0.01
7-yr	0.81 ± 0.01	0.82 ± 0.01	0.75 ± 0.02	0.80 ± 0.01	0.82 ± 0.01
Recipient survival					
1-yr	0.71 ± 0.01	0.72 ± 0.02	0.69 ± 0.01	0.72 ± 0.01	0.59 ± 0.01
3-yr	0.73 ± 0.01	0.74 ± 0.01	0.72 ± 0.01	0.73 ± 0.01	0.66 ± 0.01
5-yr	0.77 ± 0.01	0.78 ± 0.01	0.75 ± 0.01	0.76 ± 0.01	0.75 ± 0.01
7-yr	0.81 ± 0.01	0.82 ± 0.01	0.78 ± 0.01	0.80 ± 0.02	0.82 ± 0.02

Note. Prediction models presented as mean ± standard deviation.
* Proportional hazard assumption was violated in these models.

Table 3
Hosmer–Lemeshow chi-square

Time point	Logistic regression	Single-output ANN	Cox model* (no time-varying effects)	Cox model (with time-varying effects)	Multiple-output ANN
Graft survival					
1-yr	10.9 ± 5.2	12.0 ± 4.3	205.8 ± 67.0	40.7 ± 14.5	60845.6 ± 13494.1
3-yr	12.1 ± 6.1	12.0 ± 4.9	397.5 ± 129.9	34.2 ± 13.7	52014.0 ± 5849.6
5-yr	10.2 ± 2.6	11.8 ± 6.9	899.8 ± 202.4	29.3 ± 13.9	20317.8 ± 2037.5
7-yr	12.2 ± 5.9	11.1 ± 3.9	1768.9 ± 286.7	178.9 ± 29.2	16.3 ± 8.2
Recipient survival					
1-yr	11.1 ± 4.7	12.4 ± 5.8	58.9 ± 18.3	40.1 ± 9.3	33340.7 ± 4886.0
3-yr	14.1 ± 5.6	10.5 ± 5.0	63.6 ± 20.6	56.9 ± 19.2	34968.3 ± 1976.7
5-yr	11.7 ± 3.7	10.0 ± 4.6	70.5 ± 39.4	54.1 ± 10.3	13673.3 ± 626.7
7-yr	12.9 ± 6.2	12.1 ± 5.9	173.4 ± 50.7	148.8 ± 29.7	12.5 ± 4.7

Note. Prediction models presented as mean ± standard deviation.
* Proportional hazard assumption was violated in these models.

Table 4
Percentage of non-monotonic prediction

Survival variable	Logistic regression (%)	Single-output ANN (%)	Multiple-output ANN (%)
Graft survival	0.52	2.34	5.46
Recipient survival	0.24	2.81	8.40

Table 5
Agreement of the predictors' effects on short-term versus long-term survival in logistic regression models

Number of predictors		7-yr	
		Significant	Not significant
Graft survival			
1- yr	Significant	26	11
	Not significant	10	24
Recipient survival			
1- yr	Significant	16	8
	Not significant	16	31

censored at an early time point (e.g., 1-yr) is *always* censored at a later time point (e.g., 7-yr). Therefore, training samples for multiple-output ANNs were only those observations that were *not* censored at the latest time point being analyzed no matter which time points the models were predicting. In this study, multiple-output ANNs analyzed only 10,250 observations for graft survival and 11,543 for recipient survival. In contrast, logistic regression and single-output ANNs also included observations that were censored at 7-yr but not censored at 1-yr when they were making 1-yr prediction. They analyzed totally 46,414 for graft and 48,926 for recipient survival. It was not surprising that multiple-output ANNs were not able to achieve comparable performance at earlier time points since their sample size was only one fourth of all available observations. On the other hand, the three types of models had exactly the same sample size in 7-yr prediction and thus showed similar AUC measures.

We further examined the performance of logistic regression and Cox models given the same training dataset that multiple-ANNs used, namely 10,250 observations for graft survival and 11,543 for recipient survival. The AUC of logistic regression and Cox models at 1-yr, 3-yr, and 5-yr survival dropped significantly and were close to multiple-output ANNs with differences smaller than one percent. This result suggests that it is not the topology of multiple-output ANNs itself but the sample size they are able to analyze that limits performance of this type of model.

In the comparison between logistic regression and Cox models, Cox models with time-varying effects showed similar performance as logistic regression in terms of AUC measure, but Cox models without time-varying effects performed much worse. This is because the models without time-varying effects assume that the hazard ratio is constant over time [25]. That is, the effect of the predictors is the same for short-term and long-term survival. The patient-specific survival rate is therefore always in proportion to the baseline survivor function at a fixed ratio. However, this assumption may not hold for a disease like renal failure, where comorbidities such as hypertension and diabetes may progress over time and/or combine in a non-linear fashion [26]; or for a domain like transplantation, where survival depends on both donor and recipient variables that may have relatively different effects on short-term versus long-term survival.

We further examined the effect of predictors in the logistic regression models that predict short-term (i.e., 1-yr) survival versus long-term (i.e., 7-yr) survival. For either graft or recipient survival, the effect of predictors did not agree well for the short-term versus long-term survival. Several donor variables (e.g., donor type and cause of death) and transplantation parameters (e.g., cold storage time) had significant effects on short-term survival but not on long-term survival. In contrast, some recipient variables, such as history of diabetes and cause of ESRD, played important roles in long-term survival but not in short-term survival. In other words, the effects of predictors vary over time, and the proportional hazard assumption does not hold.

Once the proportional hazard assumption is violated, the model is not appropriate, and the results from the model will be misleading. Therefore, when building Cox models, it is crucial to test the proportional hazard assumption and to identify the predictors that violate the assumption. Additional efforts must be made to model the time-varying effects of such predictors in order to obtain appropriate Cox models.

In contrast, the other three types of models do not have such an assumption and are able to model the time-varying effects by their nature. In single time-point models (i.e., logistic regression and single-output ANNs), independent models are developed to predict survival at different time points and thus are able to accommodate the different effects of the same predictor in short-term versus long-term survival. Multiple-output ANNs are also able to model time-varying effects by adjusting the weights associated with the units in the output layer.

Cautions should be made when comparing models that make use of censored data (such as Cox models) and models that ignore censored data (such as logistic regression and ANNs). Predictions made by models ignoring censored data may be biased and may not be generalizable to future data [27].

5.3. Model calibration

Both logistic regression and single-output ANNs showed fairly good calibration. The calibration of Cox models was not satisfying.

Table 6
Summary of model characteristics

	Single-time-point models		Multiple-time-point models	
	Logistic regression	Single-output ANN	Cox model	Multiple-output ANN
Performance discrimination	May perform better than multiple time-point models due to ability to model different effect of predictors on short-term versus long-term survival	1. May perform better than multiple time-point models due to ability to model different effect of predictors on short-term versus long-term survival 2. May perform better than logistic regression to model complex interaction and non-linear relationships	Must identify predictors that violate proportional hazard and model time-varying effects of these predictors explicitly in order to accommodate different effect of predictors on short-term versus long-term survival	May not be satisfying due to inability to handle censored data
Calibration	May achieve satisfying performance	May achieve satisfying performance	Sensitive to baseline survival rates being specified	May not be satisfying due to inability to handle censored data
Prediction monotonicity	May generate non-monotonic prediction	May generate non-monotonic prediction	Always monotonic	1. Not able to eliminate non-monotonicity by internal mechanism 2. May be serious due to inability to handle censored data
Censored data	Partially account for censored data	Partially account for censored data	Completely account for censored data	Not able to use censored data
Implementation				
Training time	Shortest	Much longer	Short; much longer when modeling time-varying effects	Much longer
Model tuning	Limited effort required	Substantial fine-tuning required for a larger number of parameters	Limited effort required	Substantial fine-tuning required for a larger number of parameters
Interpretability	Interpretable, may be used to identify individual predictor effects	Hard to interpret, not feasible to identify individual predictor effects	Interpretable, may be used in identifying individual predictors' effect	Hard to interpret, not feasible to identify individual predictor effects

This may be because the specified baseline survivor functions were not appropriate. The predicted survival rate for a recipient depends on two factors: the hazard ratio for that particular recipient and the underlying baseline survivor function. Cox models estimate the hazard ratio and eliminate the necessity of specifying the baseline. Therefore in most studies, Cox models are used only to identify the effect of individual predictors since specifying an appropriate baseline is not trivial. However, in a prediction model, a baseline must be specified in order to calculate the recipient-specific survival rates. An inappropriate baseline may result in poor model calibration but does not affect the AUC measurement because AUC measurement depends on the order of the predicted rates but not the exact values of the rates [11].

Model calibrations of multiple-output ANNs were even worse than Cox models in 1-yr to 5-yr prediction but fairly acceptable in 7-yr prediction. This might have been caused by the small sample size due to the limitation that this type of model excludes observations that were censored at any of the time points being analyzed.

Model calibration assesses a different aspect of prediction than AUC measurement. In clinical decision support, physicians should consider whether they want to provide the patient with the exact predicted survival rates (e.g., 82% survival rate at 3-yr) or a relative percentile in survival expectation (e.g., more likely to survive at 3-yr than 75% of all other patients). If it is the former scenario, good model calibration is required; while in the latter, achieving good AUC measurement will be enough. In the study results, logistic regression and single-output ANNs potentially can provide useful decision support in both scenarios whereas Cox models and multiple-output ANNs are not feasible in the second scenario. For Cox models used in the first scenario, an appropriate baseline survivor function must be specified whereas the baseline may be arbitrary or even eliminated in the second scenario.

6. Conclusion

This study compared single and multiple time-point models to predict survival in kidney transplantation. Four types of models were examined: (1) the regression version of single time-point model, logistic regression, (2) the ANN version of single time-point model, single-output ANNs, (3) the regression version of multiple time-point model, Cox model, and (4) the ANN version of multiple time-point model, multiple-output ANNs. Model performance, handling of censored data, model implementation, and model interpretability were examined and compared. Table 6 summarizes the characteristics of the models. These findings may be generalizable to survival prediction and decision support in clinical domains sharing similar characteristics with kidney transplantation from the modeling point of view.

For survival prediction, this study suggests that:

1. Single time-point and multiple time-point models can achieve comparable AUC, except for multiple-output ANNs, which may perform substantially worse when a large proportion of observations are censored at the last time point being analyzed.
2. Logistic regression is able to achieve comparable performance as ANNs when there are no strong interactions among the predictors in the dataset and when the relationships between the outcomes and individual predictors are relatively linear.
3. When using Cox models in clinical domains where some predictors may have significantly different effects on short-term versus long-term survival, such predictors must be identified and their time-varying effects must be modeled explicitly.
4. Appropriate baseline survivor function should be specified for Cox models in order to achieve good model calibration, especially when the clinical decision support is designed to provide exact predicted survival rates instead of a relative percentile in survival.

This is the first study comparing these four types of models based on the same dataset from the view point of single and multiple time-point models. More studies need to be done in order to validate the generalizability of our findings.

Acknowledgments

The authors sincerely thank two anonymous reviewers for their insightful comments; Dr. Hajime Uno of Kitasato University, Tokyo and Dr. Tianxi Cai of Harvard University for their discussion regarding censored regression models; Mr. Bradley C. Baird of University of Utah and Mr. Randall J. Smout of ICOR for assistance in SAS and data cleaning. The data reported in this study have been supplied by the USRDS. The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as official policy or interpretation of the U.S. Government.

References

[1] Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform 2001;34(6):428–39.
[2] Abbott RD. Logistic regression in survival analysis. Am J Epidemiol 1985;121(3):465–71.
[3] Cox DR. Analysis of survival data. London: Chapman & Hall; 1984.
[4] Grambsch P, Therneau T. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515–26.
[5] Allison P. Survival analysis using SAS: a practical guide. SAS Institute Press; 1995.
[6] Kaplan E, Meier P. Nonparametric estimation from incomplete observations. Am Stat Assoc J 1958;15:457–81.
[7] Siddiqi N, McBride MA, Hariharan S. Similar risk profiles for post-transplant renal dysfunction and long-term graft failure: UNOS/OPTN database analysis. Kidney Int 2004;65(5):1906–13.
[8] Goldfarb-Rumyantzev AS, Pappas L, Scandling JD, Smout RJ, Horn S. Prediction of three year cadaveric graft survival based on pre-transplant variables in a large national dataset. Clin Transpl 2003;17(6):485–97.
[9] Mitchell T. Machine learning. McGraw-Hill; 1997.
[10] Moller M. A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw 1993;6:525–33.
[11] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35(5-6):352–9.
[12] Ohno-Machado L, Musen MA. Sequential versus standard neural networks for pattern recognition: an example using the domain of coronary heart disease. Comput Biol Med 1997;27(4):267–81.
[13] Ripley BD, Ripley RM. Neural networks as statistical methods in survival analysis. In: Dybowski R, Gant V, editors. Artificial neural networks: prospects for medicine. Landes Biosciences Publishers; 1998.
[14] Furness PN, Levesley J, Luo Z, Taub N, Kazi JI, Bates WD, et al. A neural network approach to the biopsy diagnosis of early acute renal transplant rejection. Histopathology 1999;35(5):461–7.
[15] Brier ME, Ray PC, Klein JB. Prediction of delayed renal allograft function using an artificial neural network. Nephrol Dial Transpl 2003;18(12):2655–9.
[16] Camps-Valls G, Porta-Oltra B, Soria-Olivas E, Martin-Guerrero JD, Serrano-Lopez AJ, Perez-Ruixo JJ, et al. Prediction of cyclosporine dosage in patients after kidney transplantation using neural networks. IEEE Trans Biomed Eng 2003;50(4):442–8.
[17] Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer 2001;91(8 Suppl):1636–42.
[18] Ando T, Suguro M, Kobayashi T, Seto M, Honda H. Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling. Cancer Sci 2003;94(10):906–13.
[19] Hatzakis GE, Tsoukas CM. Neural networks morbidity and mortality modeling during loss of HIV T-cell homeostasis. Proc AMIA Symp 2002:320–4.
[20] Ishitani M, Isaacs R, Norwood V, Nock S, Lobo P. Predictors of graft survival in pediatric living-related kidney transplant recipients. Transplantation 2000;70(2):288–92.
[21] Gjertson DW, Cecka JM. Determinants of long-term survival of pediatric kidney grafts reported to the united network for organ sharing kidney transplant registry. Pediatr Transpl 2001;5(1):5–15.

[22] Shortliffe SH, Perreault LE, Wiederhold G, Fagan LM. Medical informatics: computer applications in health care and biomedicine. 2nd ed. Springer; 2000.
[23] Hosmer DW, Lemeshow S. Applied logistic regression. Wiley; 1989.
[24] Mock P. Empirical comparisons of proportional hazards and logistic regression models. Stat Med 1990;9(4):463–4.
[25] Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med 2002;21(15):2175–97.
[26] Cheung AK, Agodoa L, Daugirdas JT, Greene T, Levey AS, Milford E, et al. Predictive value of blood pressure for mortality in chronic hemodialysis patients changes with duration of follow-up. J Am Soc Nephrol 2003;14:2A.
[27] Uno H, Cai T, Tian L, Wei L. Evaluating prediction rules for *t*-year survivors with censored regression models. J Am Stat Assoc 2007;102(478):527–37.